<기말 리포트>

먼저 Training set과 Test set 데이터를 읽었다.

Dim 과 summary를 이용해 데이터의 특징을 알아보려고 했는데 na가 많다는 점과 데이터가 매우 커서 전처리 과정이 매우 복잡할 것이라는 것을 알게 되었다. 45028개의 데이터와 231개의 variable로 이루어진 데이터였고summary와 dim을 통해서 알 수 있는 것들이 데이터가 너무 많아서 한계가 있었다.

NA에 대해서 다시 알아보기 위해 ColSums(is.na(train))을 활용하였는데 이때 대부분의 데이터가 NA임을 다시 한번 알게 되었다. 대부분의 variable들이 40000개 이상의 데이터를 NA로 가지고 있기 때문에 이를 다 지울 수는 없었다. 다 지우는 순간 데이터의 대부분이 날라가기에 na 값들에 각 열의 평균 값들을 넣어주었다. 하지만 이때 간과한 점이 factor값들을 지닌 열들은 na가 존재 하더라도 이를 평균값으로 대체 하지 못한다는 점이다. 그렇다고 임의의 값으로 넣어주는 것도 분석의 질을 떨어뜨릴 것이라 생각하여 아예 이 열들을 지워 주기로 했다. 따라서 다시 train\_n와 test\_n를 만들어 na를 평균값으로 대체 하지 못하는 열들을 다 지웠다.

NA 값들에 대해서 모든 처리를 하였고, 이제는 calibration set을 만들었다. Test set에 churn값이 없기에 주어진 training set을 쪼개 하나는 모델을 만드는 데 사용하고 나머지 하나는 모델의 성능을 검증하는데 쓰기 위함이다. 이때 데이터가 랜덤하게 나누기 위하여 runif함수를 활용했다.

이제 데이터의 전 처리는 어느 정도 끝이 났다. Churn 데이터를 예측하기 위한 최고의 모델을 선택하고 이를 최적화 하는 과정이 남았다. 우리가 배운 여러 모델 중 어느 것을 활용할까 생각해보았는데 logistic regression이 최고의 모델로 판단하였다. 그 이유는 다음과 같다. 먼저 아까 데이터를 가공한 후 남은 variable는 193개였다. 그 중 numeric value가 173개였다. 또한 우리가 해결해야 할 문제는 binary classification문제 였다. 대부분의 데이터 타입이 numeric인 상황에서 binary classification을 나눌 때는 logistic regression이 효과적이다. Numeric value들이 약간씩 변할 때 마다 시그노이드 함수 값들이 조금씩 변하면서 확률이 계산되 가장 정확한 확률을 구할 수 있다고 판단하였다. 따라서 모델을 logistic regression으로 선정하였고, 이를 위해서 churn 데이터의 -1값을 0으로 바꾸어 주었다.(시그노이드 함수는 0,1을 분류 값으로 가져야하기 때문에)

이제 모델의 최적화를 위해 어떤 변수를 넣어줄까 고민하였다. 처음에는 모든 변수를 다 넣어 봤는데 에러가 걸렸다. factor 변수들이 들어가면서 에러가 생기는 것을 발견하였다. 예를 들어 Var 228의 경우 Train data와 Test data의 levels가 달랐다. 따라서 factor 변수들을 포함할 경우 모델이 잘 만들어지지 않아 펙터 변수들을 제외하고 모델을 만들어야 겠다고 판단하였다. 데이터를 다시 보니 많은 데이터들이 거의 같은 값들을 가지고 있었다. 즉 Min, 1st Qu ,Mean, 3rd Qu 값들이 거의 같았다. 이러한 변수들은 churn을 분류하는데 있어서 크게 영향을 미치지 않을 것이라 판단하여 제외하고 데이터들이 실제 차이가 많이 나는 variable들을 골라 모델을 만들었다. 그 결과 0.59~0.62정도의 auc를 Calibration데이터에서 얻을 수 있었다.

한가지 개선할 점은 모델에 의미 있는 데이터 들만 넣으면 더 좋은 모델을 만들 수 있다는 생각이었다. Numeric 변수들의 선별 과정을 거치지 않고 모든 numeric 변수들을 넣어보았다. 이때 auc가 0.66정도 나오게 되었다. 즉 눈에 보일 때는 의미 없다고 생각했던 변수들도 실제 모델 안에 들어가면 성능을 많이 향상 시켜 준다는 사실을 시험이 끝나고 알게 되었다. 안타깝지만 좋은 배움 이었고 펙터 변수를 정제해 넣어주게 된다면 더 성능이 향상될 것으로 기대된다. 하지만 테스트 모델에서 이를 예측할 때 계속 오류가 생겨 이 방법을 시험시간에는 포기하게 되었다.

../../../../../Desktop/Screen%20Shot%202017-06-13%20at%203.39.0